

A Taxonomy and Survey on eScience as a Service in the Cloud

Amelie Chi Zhou, Bingsheng He and Shadi Ibrahim

Abstract—Cloud computing has recently evolved as a popular computing infrastructure for many applications. Scientific computing, which was mainly hosted in private clusters and grids, has started to migrate development and deployment to the public cloud environment. eScience as a service becomes an emerging and promising direction for science computing. We review recent efforts in developing and deploying scientific computing applications in the cloud. In particular, we introduce a taxonomy specifically designed for scientific computing in the cloud, and further review the taxonomy with four major kinds of science applications, including life sciences, physics sciences, social and humanities sciences, and climate and earth sciences. Our major finding is that, despite existing efforts in developing cloud-based eScience, eScience still has a long way to go to fully unlock the power of cloud computing paradigm. Therefore, we present the challenges and opportunities in the future development of cloud-based eScience services, and call for collaborations and innovations from both the scientific and computer system communities to address those challenges.



1 INTRODUCTION

The development of computer science and technology widens our view to the world. As a result, the amount of data observed from the world to be stored and processed has also become larger. Analysis of such large-scale data with traditional technologies would be too time consuming to hinder the development of scientific discoveries and theories. eScience is the kind of science specifically proposed to address large-scale data problems. It is the tool that offers scientists the scope to store, interpret, analyze and distribute their data to other research groups. eScience will play a significant role in every aspect of scientific research, starting from the initial theory-based research through simulations, systematical testing and verification to the organized collecting, processing and interpretation of scientific data. Recently, cloud computing has been considered as the computing infrastructure for eScience. This survey paper reviews the status of cloud-based eScience and further identifies the challenges and opportunities along this line of research.

Although the term of eScience has only been used for about a decade, the study of eScience problems started much earlier. In the early days, scientists from various fields couldn't really capture, organize and analyze the large-scale scientific data, hindering the development of science. Technological advances such as the computer and Internet have brought eScience study to a new stage. eScience projects in various fields such as biology, chemistry, physics and sociology are emerging [1], [2], [3], [4], benefiting from the platforms and toolkits in computer science and development experience shared by other research groups in domain fields. Grid computing has greatly advanced the development of eScience. Currently, almost all major eScience projects are hosted in the grid or cluster

environments [5]. With aggregated computational power and storage capacity, grids are able to host the vast amount of data generated by eScience applications and efficiently conduct data analysis. This has enabled researchers to collaboratively work with other professionals around the world and to handle data enormously larger in size than before. Many countries have devoted much investment to build their own grid platform, such as GridPP [5] in the UK and TeraGrid in US, CNGrid in China, and so on.

In the last few years, the emergence of cloud computing has brought the development of eScience to another new stage. Cloud computing has the advantages of scalability, high capacity and easy accessibility compared to grids. Recently, many eScience projects from various research areas have been shifting to cloud platforms [6], [7], [8]. eScience as a service becomes an emerging and promising direction for science computing. This survey focuses on the cloud services and techniques adopted in current eScience projects from the infrastructure, ownership, application, processing tools, storage, security, service models and collaboration aspects.

The service model and well-developed tools in the cloud platform have offered great opportunities for eScience research. The service model of the cloud relieves the users from the low-level infrastructure problems. Cloud resources are easy accessible, which makes it possible for researchers in small organizations to deal with large-scale data. The well-developed tools in the cloud, including workflow systems such as DAGMan [9] and new cloud oriented programming models such as MapReduce and DryadLINQ greatly reduce the development cycle of the eScience projects and the risk of development faults as well. People from database community are building scientific databases such as SciDB to better fit the requirements of eScience. Various experiments with eScience projects conducted on both the cloud and clusters are revealing the benefits of doing science on the cloud, helping researchers to make their choices.

While offering new development opportunities for eScience, the cloud platform also introduces new challenges for devel-

• A-C. Zhou and B-S. He, School of Computer Engineering, Nanyang Technological University, Singapore.

• S. Ibrahim, INRIA Rennes - Bretagne Atlantique, Rennes, France.

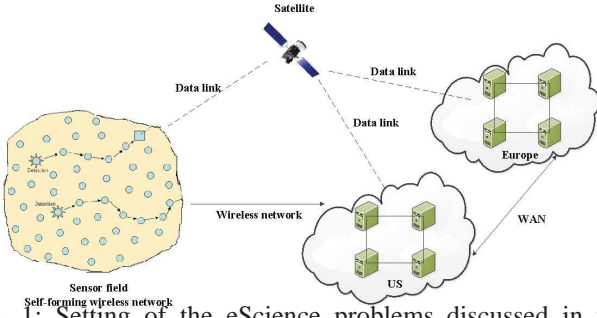


Fig. 1: Setting of the eScience problems discussed in this article

oping eScience services. Due to the pay-as-you-go pricing model, users of the cloud need to properly plan their execution, as it is not trivial to minimize the cost. Furthermore, the easy accessibility and resource sharing mechanism of cloud computing introduces security issues around storing sensitive data in the cloud. In order to ensure the confidentiality of their data from other cloud users, they need to design their own security mechanism and implement them on the cloud. A cloud platform also has the problem of data lock-in, because the current cloud providers do not have standardization on the services they provide. Thus, moving data from one cloud to another is not trivial. All these challenges require hard work and close collaboration between domain experts in computer science and eScience.

Although previous work has surveyed eScience and cloud computing separately, few of them have provided a review from the point of view of eScience in the cloud [10], [11]. Both eScience and cloud computing are rapidly developing and becoming more mature. It is timely to examine the efforts and future work for scientific computing in the cloud. This article focuses especially on eScience projects in the cloud and comparing the advantages and weaknesses with eScience in the grid to discuss the obstacles and opportunities of eScience services in the cloud.

The rest of the article is structured as follows. Section 2 introduces the background information of eScience history, grid-based eScience and cloud-based eScience. Section 3 gives the taxonomy of eScience in the cloud. Section 4 presents some eScience example projects on the cloud from four different scientific areas. Section 5 discusses about the obstacles and opportunities for eScience projects on the cloud. Finally, Section 6 draws conclusions from the article.

2 BACKGROUND

In this section, we briefly discuss some history remarks on scientific computing development, particularly for eScience. Next, we focus our review on the grid based scientific computing, and introduce the background for cloud computing.

2.1 History Remarks

Due to intensive computational and data requirements from scientific computing, computer infrastructures have been adopted to host scientific data sets and computations. eScience is a new science paradigm that uses distributed computing

TABLE 1: Development stages of the scientific computing.

Stage	Data Generated	Research Period	Infor. Tech.
Manual	By hand	Ad-hoc	Paper and pencil
(Semi-) Automated	With the help of machinery	Short-term	Computer assisted
Large-scale Sensing	From satellites and sensors around the world	Real-time	Cluster and grid

infrastructures for computation, simulation and analysis. In addition, the scientists can make use of high speed network to access huge distributed and shared data collected by sensors or stored in database. This distributed HPC and data environment allows scientists around the world to share knowledge and resources, and build close scientific collaborations.

The term *eScience* was first proposed in 1999 and was further interpreted by more researchers since then [12]. During the development of eScience, we believe it has gone through several stages to evolve from traditional science to the eScience today. Table 1 shows the major development stages the scientific computing has gone through. We review the history in the following dimensions.

Dimension 1: the evolution of science. We observed that technology (particularly information technology) is one of the main driving factors in pushing science forward. From the perspective of experimental methods, eScience first used *manual* measurements: meaning the measurements were taken by hand, not using machinery or electronics to fulfill the function. Then with the development of technology, machinery such as computers and metering instruments are used to help in the measurements, but with manual operations still involved. This stage is called the *semi-automated* stage. After this stage, machinery took a greater part in the measurements and eScience has evolved to the *automated* stage where machines took almost all the work with the least of human involvement. To recent years, new technologies such as high performance computers, sensor networks and various experimental softwares make the eScience measurements evolve to the *large-scale sensing* stage [13]. Take the research in Meteorology for example, in the early stage (classified to manual stage), researchers use thermometer, barometer, hygrometer and etc to measure the meteorological variables such as temperature, air pressure, water vapor and write down the records. They archive those meteorological data for drawing climatic maps and studying the climate of local area.

In the 19th century (classified to semi-automated stage), breakthroughs occur in meteorology after observing the development of networks. The meteorological data collected in local meteorological observatories are transmitted through networks and then are gathered together by different spatial scales to study the various meteorological phenomena.

Since the 20th century (classified to automated stage), with the adoption of radars, lasers, remote sensors and satellites into the meteorological research, collecting data of a large area is no longer a challenging problem and special instruments together with the automation of computers can automatically fulfill the measuring tasks. During this time, computers are used for doing data analysis and transmitting results for sharing. At the end of the 20th century (classified to large-

scale sensing stage), large scale observation experiments are performed. Such as during December 1977 to November 1979, back then a large scale atmospheric measurement experiment took place involving more than 100 countries around the world. This experiment was relied on satellites, meteorological rockets, meteorological observatories on the ground around the world, automatic meteorological stations, airplanes, ships, buoy stations and constant level balloons. These instruments were combined to form a complete observing system to automatically measure the meteorological variables world-wide.

Dimension 2: the length of research period. eScience has gone through *ad-hoc* stage when research was done just for a specific problem or task, and not for other general purposes later; *short-term plan* stage when researchers made plans in priori for their problems about what to do in what time, so that a project of a short term could be kept on schedule; and *real-time* stage when the research is subject to real-time constraints, such as the experimental data are collected in real-time and the system needs to give out results also in real-time. This evolution on research period also require the experimental methods to be more efficient, and the support of high technology as we will discuss next.

Dimension 3: the technology. eScience has gone through *paper and pencil* stage when no machinery was involved in our research and human work with paper and pencil was the only tool for science; then computers appeared and eScience was thus able to move to the *computer assisted* stage when computers played a great role in helping with complex calculations and solving logical problems; with the scientific problems getting more complicated and traditional computers not sufficient for the computing power required, *cluster and grid* are coming to scientists' vision and help them solving many data-intensive or compute-intensive problems within reasonable time which is not possible on traditional computers.

We summarize our findings in the three dimensions. Scientists only deal with specific problems using manual methods such as doing theoretical calculation using paper and pencil at early days. As problems getting more complicated, more planning is needed for the research and semi-automated and automated methods are also required in the research during this time. Computers are used and when problem scale gets larger, new technologies such as clusters and grids are applied for solving the problems faster. What's the next step? When problem scales get even larger and the big data coming into sight, also with the real-time constraints on the problems, even clusters and grids are not enough to tackle such problems. Recently, many eScience projects are leveraging the technology of cloud computing [1], [4], [14], [15]. With its high performance, scalable and easy accessible characteristics, it will offer new opportunities for the new problems.

2.2 Grid-based eScience

Current major eScience projects are mostly hosted in the grid or HPC cluster environment. With aggregated computational power and storage capacity, grids have been considered the ideal candidate for scientific computing. There are many labs

around the world working on grid based projects, such as GridPP in UK, TeraGrid in US, CNGrid in China, France Grilles in France, D-Grid in Germany, Kidney-Grid in Australia, etc.

In UK, particle physicists and computer scientists have been collaboratively working on the GridPP project. They manage and maintain a distributed computing grid across the UK with the primary aim of providing resources to particle physicists working on the Large Hadron Collider experiments at CERN [5]. The collaboration incorporates computing facilities at the Rutherford Appleton Laboratory along with four other grid organizations of ScotGrid, NorthGrid, SouthGrid and LondonGrid. These organizations include all of the UK universities and institutions that are working as members of this project. At the end of 2011, the project has contributed a large number of resources (29,000 CPUs and 25 Petabytes of storage) to the worldwide grid infrastructure.

The Grid Infrastructure Group (GIG) along with eleven resource provider sites in the United States have initiated an eScience grid computing project called TeraGrid. TeraGrid provides high-performance computation resources, data resources and tools, and high-end experimental facilities to users all around the USA through high-performance network connections. For example, in 2007, the resources TeraGrid provided included more than 250 Teraflops of computation resources and more than 30 Petabytes of data storage resources. Researchers could access more than 100 databases of different disciplines. In late 2009, TeraGrid resources had grown to 2 Petaflops of computing capability and more than 60 Petabytes storage. In mid-2009, US National Science Foundation (NSF) extended the operation of TeraGrid to 2011.

China National Grid (CNGrid) has quickly grown to serve more than 1400 users including both research institutes and commercial companies, providing more than 380 Teraflops of computation resources and more than 2 Petabytes of shared data storage resources. Since 2009, this project has built three Petaflop-level supercomputers, in which Tianhe-1 was ranked the fastest supercomputer in the top 500 supercomputers in 2010 [16]. With the built of the three supercomputers, CNGrid resources has grown to 8 Petaflops of computation capability and supports computation services for more than 700 national research and engineering projects in the areas of meteorology, medicine and pharmacology, aircraft engineering and aerospace engineering, etc.

Another example is the Worldwide LHC Computing Grid, which involves international collaborations of more than 150 computing centers in nearly 40 countries around the world. The European Grid Infrastructure (EGI) [17], the Open Science Grid and the Nordic Data Grid Facility, etc, are all participants of this project. It consists of a grid-based computer network infrastructure to utilize the global computation resources for storing, distributing and processing the large volume of data (around 25 Petabytes per year) produced by the Large Hadron Collider (LHC) experiments. At the end of 2010, the Grid consisted of 200 thousand processing cores and 150 Petabytes of disk space, distributed across 34 countries.

Besides the collaborations between major research centers, volunteer computing projects are taken place to build

grid platforms with public donation of computing resources. SETI@home [18] is such a volunteer computing project employing the BOINC software platform to search for extraterrestrial signals with the spare capacity on home and office computers.

The strength of grid computing has attract many scientific applications to work on grids.

- First, since governments are very concerned about the research on grid and frontier scientific research, most of the grid-based projects are funded by national fundings. Such as the GridPP project is funded by the UK's Science and Technology Facilities Council with a total amount of 47 million pounds till 2011; the TeraGrid project received 98 million dollars from NSF by 2004, 150 million dollars extended support in 2005 and another 121 million in 2011; the CNGrid project received around 94 billion Chinese yuan from 2006 to 2010. Sufficient amount of money offers good chances for institutes to hire highly qualified domain experts to do research and equip powerful computers and other resources.
- Second, single research institute can enjoy the vast computational and storage resources from grids by donating their own idle resources. Such institutes may not have enough budget for them to buy powerful computers or build their own data centers.
- Third, the tools and softwares developed on grid can benefit more research groups besides the developers themselves. This strength can save a lot of development time for the projects developed on the grids.

While Grid is the dominant infrastructure for eScience, it faces a number of limitations. First, due to the development of sensors and storage techniques, many data-intensive eScience applications are emerging. Even with the powerful supercomputers, grid may no longer satisfy the need of capacity. Second, due to the limitation of its structure, grid is not able to provide the elasticity required by most scientific projects which are pursuing cost efficiency. Third, it's not easy to get access to grid resources for everyone because a program getting access to grid resources needs to be authorized on the project's behalf and resources would then be distributed to this project as a whole. Since grids are mostly national-wide initiatives, getting the authorization is very hard for most small-scale projects. Finally, while Grid offers access to many heterogeneous resources, many applications need very specific and consistent environments. Due to these reasons, many of the eScience applications are shifting to the Cloud which has elastic storage and computing power.

2.3 Cloud Computing

According to the definition of the National Institute of Science and Technology (NIST), cloud computing is

“The delivery of computing as a service rather than a product, whereby shared resources, software, and information are provided to computers and other devices as a utility (like the electricity grid) over a network (typically the

Internet)” [19].

Cloud computing hasn't come into popularity until early 2000's, when a lot of research efforts on the cloud were emerging. Officially launched in 2006, Amazon Web Service (AWS) is the first utility computing platform that provides computation resources as services to external customers. Many other cloud service providers, including Microsoft (Microsoft's Azure), Google (Google's Cloud Platform) and OpenStack, have come into the market since then. Open-source systems and research projects are developed to facilitate the use of cloud. Initially released in early 2008, Eucalyptus is an open-source system for deploying AWS-compatible private and hybrid cloud computing environments. In the same year, the OpenNebula toolkit was released, which is also designed for building private and hybrid clouds but with different design principles from Eucalyptus.

Cloud computing bares many similarities and differences with grid computing. In the year 2008, Foster et al. [20] has compared clouds and grids mainly from a technological perspective. Five years have passed, and we should take a revisit on those differences to catch up the recent rapid development of cloud computing and highlight its relevance to the requirement of eScience.

Compared to the grid, cloud has better scalability and elasticity.

- When developing applications on the grid infrastructure, it's not easy to scale up or down according to the change of data scale. But in cloud, with the use of virtualization, clients can scale up or down as they need and pay only for the resources they used.
- Virtualization techniques also increase the computation efficiency as multiple applications can be run on the same server, increase application availability since virtualization allows quick recovery from unplanned outages with no interruption in service and improves responsiveness using automated resource provisioning, monitoring and maintenance.
- Also, cloud has easier accessibility compared to grid. Users can access to commercial cloud resources through log in and use the resources as they need as long as they could pay with a credit card. In this case, even small businesses which could not afford purchasing high performance computers can also have the chance to use powerful clusters or supercomputers on their compute-intensive or data-intensive projects.

eScience applications are beginning to shift from grid to cloud platforms. The Berkeley Water Center is undertaking a series of eScience projects collaborating with Microsoft [6], [7], [8]. They utilized the Windows Azure cloud to enable rapid scientific data browsing for availability and applicability and enable environmental science via data synthesis from multiple sources. Their BWC Data Server project is developing an advanced data synthesis server. Through close interaction between computer scientists and environmental scientists, they are building new tools and approaches to benefit regional and global scale data analysis efforts [6], [7]. Another one, the California Water CyberInfrastructure

project, is developing a Water Cyberinfrastructure prototype that can be used to investigate and eventually manage water resources. In Europe, GRNET is initiating an eScience cloud in Greece [21]. GRNET is a state-owned company operating under the supervision of the Ministry of Education (General Secretariat of Research & Technology). Its main mission is to provide high-quality electronic infrastructure services to the Greek academic and research institutions. The vision of its eScience cloud is to provide virtualization and storage services for the Greek scientific community. This project starts with offering online storage of 50Gbytes for all Greek academic and research community (Pithos), then moves to provide VMs on demand and finally provide software as a service. We are also working on a eScience project based on cloud computing in Singapore [22]. The objective of our project is to leverage cloud computing techniques and sensor networks to provide real-time and large-scale monitoring and analysis for water quality. The project is aiming at providing real-time monitoring for the reservoirs based in Singapore, but the methods and models proposed could be utilized to benefit all water resources around the world. This project is funded by NRF Singapore and we are currently working on the first phase.

Since the cloud is an emerging field, many of the cloud based eScience projects are still in their early stage. This is partially because cloud computing has come to popularity only for several years and researchers haven't realized its strengths thoroughly. That motivates us to review the existing efforts on adopting cloud computing technologies to eScience, and to explore the research challenges and opportunities in that direction. Moreover, a taxonomy is useful in guiding the design and implementation of cloud-based eScience project.

Compared with the general cloud computing surveys (such as by Armbrust et al. [23]), this survey focuses on the review on the current status of eScience in the cloud, and therefore identifies the new opportunities and challenges on pushing the state-of-the-art. Our survey also goes beyond some perspective report on science cloud (for example, by Lee [24] and by Keahey [25] and Oliveira [26] in three major aspects. First, we define a taxonomy for eScience services in the cloud. To the best of our knowledge, our definition is the first taxonomy for eScience services. Second, we perform the detailed and comparative study on the existing efforts including tools, systems and projects. Second, based on the review on the existing efforts, we point out the challenges and opportunities that are close and practical as a guide for the intermediate next steps.

3 TAXONOMY OF EScience SERVICES IN THE CLOUD

The taxonomy in this section gives clear classification of cloud computing techniques used in eScience services from various perspectives, including the computation *infrastructure* for eScience applications, the *ownership* of cloud infrastructures, the eScience *application* types, the *processing tools* used for eScience applications, the *storage* model, the *security* insurance method, *service models* of the cloud and the *collaboration* goal between different research groups. Figure 2 gives a clear

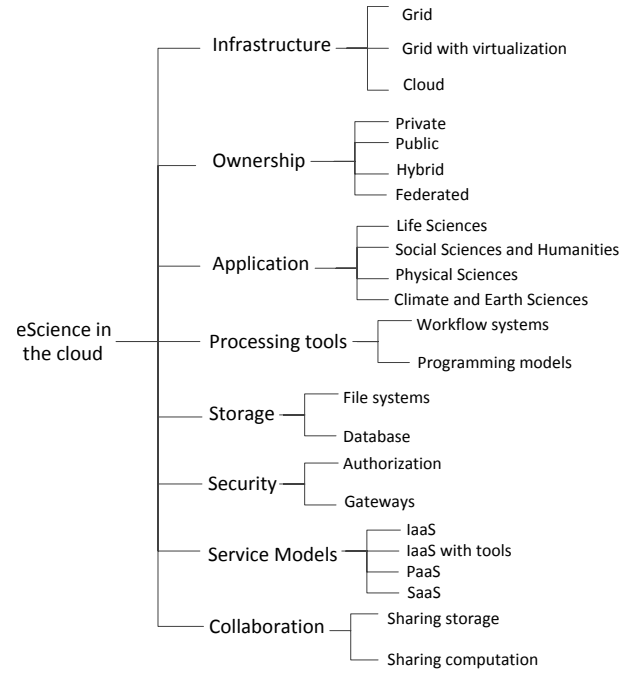


Fig. 2: Taxonomy of eScience in the Cloud

structure of the taxonomy. This taxonomy reflects the interplay between eScience and cloud computing. Some are mainly from eScience's perspective, and some are mainly from cloud computing's perspective. We introduce them one by one.

3.1 Infrastructure

The infrastructure of cloud provides access to compute and storage resources for eScience applications in an on-demand fashion. Cloud shares some similarities with Grid while at the same time is modified to overcome the limitations of Grid.

Grid computing technologies are leveraged by cloud computing to serve as its backbone and infrastructure support. Compared with grids infrastructures, cloud has pricing and monitoring services. Before 2007, most of eScience applications were implemented on Grid, where scientific organizations share their spare resources.

One characteristic of Grid is that it assigns resources to users in the unit of organizations and each individual organization holds full control of the resources assigned to it. However, this kind of resource assignment is not efficient. There are efforts in Grid to use virtualization to change this situation. Nimbus scientific cloud is one such effort that provides a virtual workspace for dynamic and secure deployment in the Grid. [27] is an astronomy application implemented using Nimbus. Virtualization hides from users the underlying infrastructures which are usually heterogeneous hardware and software resources, and provides the users with homogeneous and isolated virtual cloud environment.

In contrast to science clouds, several national cloud initiatives have also been announced to provide on-demand resources for governmental purposes [24], such as the US Cloud Storefront [28], the UK G-Cloud [29], and the Japanese Kasumigaseki [30] cloud initiatives. Many industry players also dive in the cloud business and provide users with seemingly infinite public cloud resources. With the popularity of

cloud, many eScience applications are right now transferring to the general public cloud infrastructures such as Amazon EC2, Windows Azure to benefit from its high performance, scalability and easy-access [6], [7], [8], [31], [32], [33].

3.2 Ownership

The ownership of cloud infrastructures can be classified as the following types: private, public, hybrid and federated.

Private clouds are infrastructures operated only for one single organization, no matter who the infrastructures are managed by or where they're located. The security level of private clouds is the highest among the four types. eScience applications which have high security requirements or possess highly sensitive data can be deployed on private clouds. OpenNebula is the first open-source software supporting private clouds deployment and is widely used by industry and research users right now [34]. But on the other hand, such infrastructures do not benefit from the economic models provided by the cloud since the application owners have to "buy, build and manage" the infrastructures to run their jobs.

In contrary, public clouds are more open, with their application, storage and other resources available to the public on the pay-as-you-go basis. There are quite a few commercial companies providing public cloud services, such as Amazon, Windows and Google. Many eScience applications have been deployed on this kind of cloud platforms (e.g., [32], [6], [7]) because users can easily access to the public cloud resources with a credit card.

A federated cloud, also known as community cloud, is a combination of two or more clouds from either private, public or federated clouds. In this combination, the two or more clouds often have common goals in security, compliance, jurisdiction, etc. Many countries have built federated clouds to support the research and education purpose of their own country. The EGI Federated Cloud Task Force [35] is a federation of academic private clouds to provide services for the scientific community. It has been used by a wide areas of eScience applications, including Gaia which is a global space astrometry mission [36], the Catania Science Gateway Framework (CSGF) [37] which provides science gateway for scientific application users, etc.

A hybrid cloud utilizes cloud resources from both private and public clouds. The benefit of hybrid clouds captures the best of both worlds. When the resources of the private cloud are enough to support current workload, the users will only use the private cloud to benefit from its security and stability. While the workload is bursting and the private cloud can no longer support users' requirements, users can then request resources from the public cloud to benefit from its scalability.

3.3 Application

Cloud computing techniques have been applied to various eScience applications. We have surveyed a lot of eScience papers and summarized them in the following four categories based on their areas of expertise: Life sciences [4], [14], Physical sciences [38], [27], Climate and Earth sciences [32], [6] as well as Social sciences and Humanities [39], [40].

The life sciences comprise the scientific research on living organisms, such as plants, animals, and human beings. Specifically, it includes the fields of Biochemistry, Biology, Ecology, Neuroscience, Psychology, etc. Physical sciences encompass the fields of natural science and science that study non-living systems, in contrast to the life sciences. Climate and Earth science is the study of climate and the planet Earth. The climate science is a sub-field under the atmospheric sciences which studies the average weather conditions in a period of time while the earth science includes the study of the atmosphere, hydrosphere, oceans and biosphere, as well as the solid earth. Social sciences and Humanities is the field of study concerned with society and human behaviors. It includes the scientific studies on anthropology, archeology, criminology, economics, education, linguistics, political science and international relations, sociology, geography, history, law, and psychology.

We note that those application categories can have overlaps with each other. There is no absolute boundary between each pair of categories. Still, different categories have their own requirements on the cloud. The first three categories, life sciences, physical sciences and climate and earth sciences, are more focusing on extending their works to large-scale datasets and thus require the cloud platform to deal with large-scale data analysis efficiently. The fourth category, social science and humanities, is more focusing on collaboration and thus requires the cloud platform to be easy for sharing.

3.4 Processing tools

From the perspective of processing tools, we have witnessed deployment of classic workflow systems in the cloud, new cloud oriented programming models such as MapReduce and DryadLINQ, and hybrid of such newly proposed tools and models.

Scientific workflows have been proposed and developed to assist scientists to track the evolution of their data and results. Many scientific applications use workflow systems to enable the composition and execution of complex analysis on distributed resources [41]. Montage is the example of a widely used workflow for making large-scale, science-grade images for astronomical research [27].

Workflow management systems (WMSes) such as Pegasus [42] and Kepler [43] are developed to manage and schedule the execution of scientific workflows. WMSes rely on tools such as Directed Acyclic Graph Manager (DAGMan) [9] and Condor [44] to manage the task dependencies within scientific workflows, and to manage the resource acquisition from the cloud and schedule the tasks of scientific workflows to cloud resources for execution.

The appearance of cloud oriented programming models has great promotion for the development of cloud computing. MapReduce is a framework proposed by Google in 2004 [45] for processing highly distributable problems using a large number of computers. The framework is inspired by the map and reduce functional language where the map function takes in the input, partitions it into smaller sub-problems and distributes them to multiple worker machines while the

reduce function collects the processing results to all the sub-problems and combines them in some way to form the output. Users who need to parallel their codes in order to run in distributed environment only need to define their own map and reduce functions using the MapReduce framework. This makes this framework especially suitable for eScience application users who may not be experts in parallel programming. We have observed the emergence of eScience applications adopting MapReduce framework for data-intensive scientific analyses [3].

3.5 Storage

Data is centric to eScience applications and data processing is closely related to data storage. With the development of science, the hypothesis to data has evolved from empirical description stage, theoretical modelling stage, computational simulation stage to the fourth paradigm today, the data-intensive scientific discovery stage. Due to the vast data size, knowledge on the storage format of scientific data in the cloud is very important. Normally, there are two ways for data storage: data can be stored as files in file systems or in databases.

Many distributed file systems have been proposed to provide efficient and reliable access to large-scale data using clusters of commodity hardware [46], [47]. For example, Hadoop Distributed File System (HDFS) is the primary storage system used by Hadoop applications which utilize the MapReduce model for large dataset processing. It creates multiple replicas of data blocks and distributes them on compute nodes throughout a cluster to enable reliable and rapid data access. When well-designed, features of the HDFS system such as data locality and data replication can further benefit the applications running on Hadoop via locating computation close to the data [48]. OpenStack Swift [49] is a distributed storage system for unstructured data at large scale. It currently serves the largest object storage clouds, such as Rackspace Cloud Files and IBM Sftlayer Cloud. The scalable and highly efficient distributed file system models provide a promising data storage approach for data intensive eScience applications.

Although in cloud, data storage usually relies on file systems, using databases as storage has its advantages. First, it's easier to do query in a database than in file systems since the files have to be opened and closed in order to get the data stored in. Also, database as storage can ensure data integrity. Till now, the parallel capabilities and the extensibility of relational database systems (RDBMS) were successfully used in a number of computationally-intensive analytical applications. When facing eScience applications, RDBMS have shown limitations. For one thing, not all data in eScience is relational. Several classes of "NoSQL" databases have been proposed as alternatives to RDBMS to satisfy the efficiency requirement of scientific data. For example, Amazon's Dynamo [50] is a key-value store which supports storing and retrieving data by primary key. Its key-value interface makes it especially simple and cost-effective to the cloud users. Google's Bigtable [51] is a column-oriented NoSQL database which provides column-wise as well as row-wise index for data manipulation. This

distributed storage system is designed to managing large-scale structured data: "petabytes of data across thousands of commodity servers" ([51]). Cassandra [52] is another column-oriented distributed NoSQL database which provides highly available service to large amounts of structured data. HBase, a Hadoop project modeled on Bigtable, has been applied to many eScience applications such as bioinformatics domains [53]. Some array-based databases such as SciDB [54] have also been proposed to satisfy the special requirement of array-based eScience applications. SciDB is a scientific database system built from ground up and has been applied to many scientific application areas, including astronomy, earth remote sensing, environmental studies and etc [55].

3.6 Security

Security is a big issue to eScience applications, especially for those with sensitive data. On the one hand, scientists need to make sure that the sensitive data is not stolen by people with vicious intention; on the other hand, they also need to share data between scientific groups working on the same project. Thus, how to find a balance point between the two aims is a challenging problem. Currently, the security level in the cloud is not very high compared to the Grid computing platform and the common way to make sure of security in the cloud is through logging in. Many eScience applications deployed on the cloud have designed their own way of authentication and authorization to ensure security. Such as in [56], Group Authorization Manager is used to grant access permission based on user-defined access control policy. The emerging Open Authorization (OAuth2.0) protocol is used to support authorization for users to share datasets or computing resources. In [15], the Gold security infrastructure is utilized to deal with the authentication and authorization of users to keep sensitive data secure. Data owners could specify their security preferences for the security infrastructure to control role and task based access.

Unlike in Grid computing, where the authentication and authorization mechanisms are mainly based on the public key infrastructure (PKI) protocol [57], many Cloud vendors support multiple security protocols such as OAuth2.0. The adoption of the new security protocols opens up a new design space for users to define rules of accessing secured resources and sharing data. Via the authorization delegation in the security protocols, users can define rules to allow easy collaborations between geographically distributed parties without the involvement of administrators.

3.7 Service Models

There are different levels of computing services offered by the cloud (i.e., IaaS, IaaS with tools, PaaS and SaaS). The IaaS model is the most basic cloud service model, where cloud providers only offer physical infrastructures to users, in the form of virtual machines, raw storage, and so on. Amazon EC2 is such an example [32], [33], [58]. When deploying in an IaaS cloud, users have only to install operating system and application softwares as they need. In order to save users' effort of installation, platforms providing IaaS level services but with additional tools and softwares, have been

proposed. Nimbus [27] and Eucalyptus are examples of this kind. Nimbus is an open-source toolkit that aims to deliver IaaS capabilities to the scientific community. It allows users to rapidly develop custom community-specific solutions. In the PaaS model, cloud providers provide a computing platform typically equipped with operating system, programming language execution environment, database, and web server. Users of PaaS cloud can simply develop their applications on the platform without the effort and cost of buying and managing the underlying hardware and software layers. Typical examples of this type include Windows Azure, Google's App Engine. In the SaaS model, cloud providers provide a computing platform installed with application softwares. Cloud providers are in charge of the software maintenance and support. Cloud users are eased from the trouble of managing the cloud platform and can put more of their effort on application design. Notable service providers in this class include online storage services such as Dropbox and Google Drive, online education services such as Coursera.

3.8 Collaboration

Another important usage of cloud for eScience applications is to realize collaboration. In eScience, there are more and more projects involving multiple groups closely working together on the same project and those groups are sometimes spread worldwide. The collaboration between the groups includes two different focuses. First is on sharing storage, that is the sharing of scientific data and analysis results between different research groups working on the same project. Except sharing data for collaborative works, many eScience applications open their data to the public for educational purposes. Second is on sharing of computation, that is to share the idle computing resources of one group to the others such that the resource utilization rate of all the collaborating groups can be highly improved. Collaboration between these groups is very important to the success of the projects. With the development of Internet and the popularity of social networks, many works are leveraging cloud computing techniques and social network APIs to provide a collaboration platform for eScience researchers [59], [60].

4 CURRENT STATUS

We review the current status of eScience services in the cloud, and present the key observations from our survey.

4.1 An Overview

The example systems surveyed in this section may not be exhaustive, but cover many areas of eScience researches currently going on. The table below summarized the systems from their platform, scientific operations and development and classified them by their areas from life sciences, social sciences and humanities, physical sciences and climate and earth sciences. Table 2 is a categorization of the surveyed example systems using the taxonomies introduced above. In the rest of this section, we present the major observations we have found from the example systems.

4.2 Observation 1: Ad Hoc Project Developments

The development of eScience projects is ad hoc. Some applications are developed on Amazon EC2 cloud [61], some are deployed on Windows Azure [8] while some others are developed on both cloud platforms to verify their design [62]. However, it is not clearly explained why certain cloud platforms should be chosen over others in those projects.

For example, MFA [61] is a Life Science project developed with the cloud services provided by Amazon. Its aim is to investigate whether utilizing MapReduce framework is beneficial to perform simulation tasks in the area of Systems Biology. The Monte Carlo bootstrap (MCB) method, an important building block of this application, is parallelized and implemented with Amazon Elastic MapReduce (EMR). Because of the long-running characteristic of MCB simulation, the MapReduce version of MCB is wrapped with a WSRF service which is specifically designed to support long-running operations. The experiments on a 64 node Amazon MapReduce cluster and a single node implementation have shown up to 14 times performance gain, with a total cost of on-demand resources of \$11. MODIS Azure [6] is a Climate and Earth science application deployed on Windows Azure to process large scale satellite data. The system is implemented with the Azure blob storage for data repository and Azure queue services for task scheduling. However, neither of the two projects has technically explained their choice of cloud platforms. The MapReduce framework is supported by many cloud providers other than Amazon, such as Cloudera's Distribution of Hadoop (CDH), Azure HDInsight, etc. The storage and queue services are also supported by many cloud providers besides Azure. For example, Amazon provides S3 and EBS for storage and Simple Queue Service (SQS) correspondingly. To compare the performance on different cloud platforms, an Physical science project Inversion [62] deployed its application on both Amazon EC2 and Windows Azure with symmetry structures.

All these examples indicate that, current eScience application owners are not quite clear how to choose cloud platforms. The only way for them to distinguish the cloud performance differences is through redundant implementation. Due to the diverse requirements of projects in different areas, the lessons learned during the implementation of one project may be useless to projects in other research areas.

4.3 Observation 2: Common Development Softwares and Tools

Many eScience projects, especially those in the same application class, share common development softwares and tools. For example, a workflow system such as Pegasus is widely used by Physics science applications [38], [27] where the jobs involve a number of analysis steps. Many works are proposing new techniques in the cloud for scientific applications based on the workflow model [63], [64]. However, current commercial clouds do not include such scientific tools by default. Different applications owners have to redundantly deploy and

In the Life science project MassMatrix [58], the authors used the Pegasus Workflow Management System (WMS) to create parallel workflows for a database program which searches proteins and peptides from tandem mass spectrometry

TABLE 2: Taxonomy mapping to the surveyed example systems. “-” means that aspect is not specified in the project paper.

Project	Infrastructure	Ownership	Application	Processing Tools	Storage	Security	Service Model	Collaboration
CloudBLAST	Grid with virtualization	Private	LS	MR Programming Model	File System	Authorization	IaaS	Computation
RDF	Cluster	Private	LS	MR Programming Model	Database	-	-	Computation
CARMEN	Cloud	Public	LS	Workflow System	File System / Database	Authorization	SaaS	Storage / computation
MFA	Cloud	Public	LS	Workflow System / MR Programming Model	File System	Authorization	IaaS	Computation
MassMatrix	Cloud	Public	LS	Workflow System	Database	-	IaaS with tools	Computation
LS Gateway	Cloud	Private	LS	Workflow System / MR Programming Model	File System	Gateway	SaaS	Storage / computation
CloudDRN	Cloud	Public	LS	Business Software Tools	Database	Authentication / Authorization	IaaS / PaaS	Storage
SciHm	Cloud	Public	LS	Workflow System	Database	Authorization	IaaS	Computation
SciDim	Cloud	Public	LS	Workflow System	File System	-	IaaS	Computation
Montage Example	Cloud	Public	PS	Workflow System	File System	-	IaaS	Computation
Montage Comparison	Cloud	Private	PS	Workflow System	File System	-	IaaS with tools	Computation
CGL-MapReduce	Cluster	Private	PS	MR Programming Model	File System	-	-	Computation
Kepler	Grid / Cloud	Private / Public	PS	Workflow System	File System	-	IaaS	Computation
Inversion	Cloud	Public	PS	Programming Model	File System	Authorization	IaaS	Computation
CAOCM	Cloud	Public	CES	MPI Programming Model	File System	Authorization	IaaS	Computation
MODISAzure	Cloud	Public	CES	Workflow System	File System	Authorization	PaaS	Storage / computation
RPSS	Cloud	Public	CES	Multi-threading Programming Model	File System	-	PaaS	Storage / computation
NG-TEPHRA	Grid / Cloud	Private / Public	CES	Workflow System	File System	-	IaaS	Computation
Cloudbrusting	Grid / Cloud	Private / Public	CES	Workflow System	File System	-	PaaS	Computation
SLOSH	Cloud	Public	CES	Workflow System	File System	-	PaaS	Computation
FMVE	Cloud	Private	SSH	Programming Model	File System	Authorization	IaaS	Storage
IAS	Cloud	Public	SSH	-	File System	Gateway	SaaS	Storage / computation
BetterLife2.0	Cloud	Public	SSH	MR Programming Model	File System	Authorization	IaaS	Computation
SoCC	Cloud	Public	SSH	-	File System	Authorization	PaaS	Computation
SCC	Cloud	Public	SSH	-	File System	Authorization	SaaS	Storage / computation

data. DAGMan is used to manage the data dependencies in the workflow and Condor is used to schedule the workflow. Similarly, the Physical science projects Montage Comparison [27] and Kepler [42] also utilized Pegasus-WMS, DAGMan and Condor to manage the execution of astronomy workflows. In all three applications, the authors have to separately deploy and configure all the required softwares such as Pegasus and Condor on the cloud platforms to make their applications run. Such re-implementation and re-design work requires good effort from the application owners and should be avoided.

4.4 Observation 3: Static Data Storage

Data is the centric of eScience applications. Although the data size of most eScience applications is enormous, we have observed that many of the eScience data are statically stored. For example, the SciHMM [65] project is making optimizations on time and money for the phylogenetic analysis problem. The data involved in this application are genetic data, which do not require frequent update and can be viewed as statically stored. Similarly, the bioinformatics data in the CloudBLAST [4] project and the astronomy data in the Montage Example [38], although may be updated from time to time, are seldomly modified once obtained. Once such data are uploaded to the cloud, not much networking is required to modify them. This characteristic of scientific applications makes them appropriate to be implemented on the cloud since networking usually causes the most monetary cost and overhead.

4.5 Observation 4: Privacy vs. Sharing

Data privacy and security is a big issue to scientific applications. Traditional storage at Grid and private clusters provides a high security level to scientific data through authorization and authentication. However, there is an increasing need of eScience applications to collaborate and share. Such need forces them to move their applications from traditional computing platforms to the public cloud, which in turn makes the privacy issue more serious.

One example is the Life science project CloudDRN [66]. CloudDRN moves medical research data to the cloud to enable secure collaboration and sharing of distributed data sets. It relies on authentication and authorization to ensure security. Also, many applications in Social Science and Humanities have shown such a trend. The SoCC [59] project leverages social network platform for the sharing of resources in scientific communities. They provide a PaaS social cloud framework for users to share resources and support creating virtual organizations and virtual clusters for collaborating users. The SCC [60] project is also leveraging social network and cloud computing to enable sharing between social network users. But different from previous works, it argues that since online relationships in social networks are often based on the real world relationships, it can be used to infer the trust levels between users. The benefit is users can thus share data and applications with lower privacy concerns and security overheads. In both examples, the social network information is utilized to lower the privacy and security level of the applications. Different from the authorization and authentication in Grid, this is a new privacy insurance method enabled by sharing in the cloud.

4.6 Observation 5: Performance vs. Scalability

Comparison between the implementation on HPC with implementation on the cloud is always a hot topic for scientific applications.

The NG-TEPHRA [33] project performed a volcanic ash dispersion simulation on both grid and cloud, using the East Cluster at Monash University and the Amazon EC2 computing resources separately. Experiments show efficient results on both platforms and the EC2 results have shown very small differences in their standard deviation, indicating the consistent QoS of the cloud. The MODISAzure [6] project implemented its application on both Windows Azure cloud and a local high-end desktop machine. Evaluation on a single computational instance in Windows Azure compared with that in the desktop machine shows the task execution time of the Azure instance is always longer than that of the desktop machine while the communication time is not as stable as the computation time and does not show consistent results during the experiments. When using multiple Azure instances to compare with desktop machines, the performance of the pipeline scales almost lineally with the number of Azure instances. Cloudbursting [8] implemented its satellite image processing application with three different versions: an all-cloud design on Windows Azure, a version that runs in-house on Windows HPC clusters and a hybrid cloudbursting version that utilizes both in-house and cloud resources. The hybrid version achieves the best of the previous two versions, namely the development environment of a local machine and the scalability of the cloud. Their experimental results showed that the application is benefiting from the hybrid design, both on execution time and cost.

The common observation from the above examples is that the performance comparison between cloud and HPC is application dependent. Due to the scheduling and communication overhead, the applications involving large and frequent data transfer over multiple computation nodes usually perform worse on the cloud than on HPC clusters which are equipped with high bandwidth network. But the advantage of cloud is its high scalability. Users can easily and quickly scale up and down their applications as needed, without wasting too much money. Applications such as Cloudbursting [8] can benefit from this characteristic of the cloud.

4.7 Observation 6: Utilizing vs. Advancing Cloud Computing

Many projects in various research areas are trying to benefit from the advanced cloud computing techniques. However, most of the eScience projects in the cloud are simply using cloud computing techniques to improve their applications.

For example, the Climate and Earth science project SLOSH [67] studies the efficiency of several middleware alternatives for storm surge predictions in Windows Azure. The Life science project CloudBLAST [4] uses MapReduce programming model to parallelize and speedup its programs, in order to provide distributed services for bioinformatics applications. Another Life science project RDF [14] is also using MapReduce model and Hadoop implementation to speed up the querying and reasoning over large scale resource

description framework. Many projects in Social Sciences and Humanities are utilizing science gateways and social networks to enable collaboration. FMVE [39] proposes an IT model based on several existing technologies to enable accessing and hosting applications on social network for enterprises. Better-Life2.0 [2] provides intelligent reasoning for online and mobile users through social network interfaces. LS Gateway [56] builds a science gateway to facilitate the sharing between life scientists. It adopts the OAuth2.0 protocol to support authorization for users to share data or computation resources. Many other softwares and tools, such as the Pegasus-WMS mentioned above, are also utilized to parallelize eScience applications and to facilitate their execution.

A few projects dig deeper and improve the cloud computing techniques to better fit their specific applications. For example, the CGL-MapReduce project [3] proposes a new MapReduce implementation for data intensive scientific data analysis to compare with Hadoop. CGL-MapReduce uses streaming for all communications, thus eliminates the overheads in communicating via a file system.

4.8 Observation 7: Monetary Cost is a Concern

Many applications have reported their implementation on the cloud platform from the performance perspective. However, another important consideration of eScience in the cloud, the monetary cost, is only studied by a few example systems.

MFA [61] reported a 14 times speedup for their metabolic flux analysis on Amazon cloud with a \$11 cost, which includes the EC2 cost, EMR cost and S3 storage cost. SciHmm [65] aims to reduce monetary cost for scientists via deciding the most adequate scientific analysis method for the scientists a priori. It reported the cost for the parallel execution of SciHmm on the Amazon EC2 cloud and showed it's acceptable for most scientists (US \$47.79). Another project SciDim [68] aims to optimize the total execution time of scientific workflows with budget constraints through finding the best initial configuration of the cloud. Cloud users have to pay for all the resources they have used on the cloud, including computation, network and storage resources, etc. Due to the large scale of data and long running jobs, eScience applications have to carefully plan their use of cloud to optimize their monetary cost. However, this planning is not trivial and requires both domain expertise and knowledge on cloud computing.

4.9 Summary

Although the current eScience system designs are far from mature, some common trends in all of the above eScience areas have shed light on the importance of cloud to eScience: data are easier to get and data size is increasing tremendously; the need of sharing data and computation and collaboration between scientists are also increasing. Cloud computing fits in the trends perfectly. The scalability of the cloud could offer seemingly infinite storage and computing resources for eScience applications along with the increase of scientific data. Also, the easy access to the cloud resources offers great opportunity for scientists in different locations to work on the same project.

In spite of the silver lining of developing eScience applications in the cloud, there are still problems to solve, challenges to overcome. The easy access to the cloud brings the security issue, the pricing model of the cloud brings the cost-efficiency problem and the different design between different cloud platforms also brings us the lock-in problem. All in all, for the development of eScience applications in the cloud, we still have a long way to go.

5 CHALLENGES AND OPPORTUNITIES

Previous sections have reviewed the status and the observations in building eScience applications and systems in the cloud. Despite the fruitful results along this research direction, we clearly see that there are still many open problems to be addressed in order to fully unleash the power of cloud computing for eScience. In this section, we discuss several open problems, followed by the opportunities for addressing those open problems.

5.1 Open Problems

We present the open problems for developing the next-generation eScience applications and systems in the cloud. Those open problems are rooted at the interplay between eScience requirements and cloud computing features.

Data Lock-In: So far, there are no eScience applications and systems that have been deployed on multiple cloud providers. There's no standardization between different cloud platforms, such as different clouds use different data storage formats. For example, data stored in Amazon S3 cannot be easily used by the jobs running on the Windows Azure platform due to different APIs, data storage techniques such as encryption technique and security protocols. On the other hand, due to the eScience projects usually involve a large amount of data for scientific research, such as the genome sequence data and seismographic data, data transfer cost between different cloud platforms is substantial. This also makes the data lock-in problem significant to e-Scientists.

Performance Unpredictability: Some eScience applications have rather rigid performance requirements. Performance unpredictability is a critical problem for running those applications in the cloud, due to the interference among concurrent applications running in the same cloud. This problem is particularly severe for disk I/O and network traffic. For eScience applications, this problem is especially prominent since there are a lot of read tasks needed to get input data and parameters from local disks to do data analysis and also a lot of write tasks to save the intermediate analysis results to local disks. The other factor of performance unpredictability is VM failures or unreliability. In [7], the authors issued a total of 10032 VM unique instance start events on Windows Azure cloud and only 8568 instances started once during their lifetimes while the others had encountered various unknown problems during their run and were restarted by the Azure infrastructure for many times.

Data Confidentiality and Auditability: Current commercial clouds are essentially open to public and are consequently exposing themselves to more attacks. Safety is the biggest

concern that prevents customers from storing their sensitive corporate data in the cloud. Especially for eScience applications, the data involved could be relevant to the homeland security of a country, such as the geographical data of the country, or even the security of human beings, such as the human genome data. So protecting these sensitive data from unauthorized or even malicious access is an important ongoing research topic.

Debugging: Bugs in Large Distributed Systems cannot be reproduced in smaller configurations. Although many eScience programs have been tested and evaluated in the grid and cluster environments, program debugging and testing are still challenging in the cloud.

Lacking of eScience Common System Infrastructure. As we discussed in the previous section, the efforts of implementing eScience projects on the cloud are quite ad-hoc. The effort for one project is usually not reusable for other projects. For example, the data processing softwares and interface APIs used in different scientific areas are quite different. In physical sciences, the Montage workflow, an astronomy toolkit, is commonly used to discuss the pros and cons of using cloud computing for scientific applications [38], [27] and such physical science systems built in the cloud are specifically designed to better fit the cloud for scientific workflow applications. Thus, such developmental experiences may not be useful to scientific applications in other areas, such as social sciences and humanities in which resource sharing is much more the concern and social network APIs are needed to build the social science systems. Since current eScience systems are specifically designed for each project, new projects coming into the cloud have to build their systems from top down. In order to save the development cycle and better exploit the experiences of current systems, we need a holistic platform which applications from various research fields can build their systems upon and offers opportunities for application specific optimizations.

5.2 Opportunities

We also see some opportunities in addressing those open problems. Many of those opportunities are driven by different communities outside scientists, including open-source software developers, system researchers and governments.

Open-source Cloud Software Stacks: With the popularity of cloud computing, there are a lot of cloud platforms with various architectures open to the public. Public clouds such as Amazon EC2, Windows Azure and Google App Engine own and operate cloud infrastructure and offer access to the public via Internet. Private clouds implemented using software platforms such as Eucalyptus and Nimbus on computer clusters provide hosted services to a limited number of users behind a firewall. Private cloud is operated for a single organization only, whether hosted and managed internally or externally. Hybrid cloud is the composition of two or more clouds, either private or public, to capture the best of both worlds: ability to immediately deliver services that users demand independent of Internet connectivity as well as the scalability to handle cloudbursting, an instant spike in demand. Examples of hybrid cloud include Intel Hybrid Cloud Program [69] and GoGrid

Cloud Hosting [70]. Given an application, how to choose the most appropriate cloud platform from the various kinds of cloud platforms is a very challenging issue. To solve this problem, one has to consider the characters of the application itself, such as whether it's data-intensive or compute-intensive, whether fault tolerance is important to this application, etc; consider the demand of this system, such as whether the computation demand is stable or may have instant spike of workload; also consider the aim of sharing, for example, if the aim is to share data and resources between limited users or the general public.

Towards Common System Infrastructure Support for eScience: Researchers from database community are building scientific database which can better fit the requirements of scientific applications. SciDB is such an example. In March 2008, the first SciDB workshop was held in Asilomar and representatives from both scientific community and database research community participated in this workshop. One major result of this workshop was a set of requirements that a database management system should meet in order to support the storage and analysis of several fields of data-intensive science over the next decade [71]. According to these requirements, the SciDB should provide several new features such as direct support for arrays as a first-class column type because all sciences need to work with non-scalar values like vectors and arrays, association of data element with "error bar" because all sciences must deal with observations and derived data that have inherent uncertainties, etc. The SciDB developers meeting and Open SciDB community meeting were held between 2008 to 2011 when SciDB was eventually built up and tested. The overview of SciDB was presented at SIGMOD 2010 [54] and caught a lot of attention from both scientific community and database community. There have been several use cases from various sciences for SciDB including Optical astronomy, Radio astronomy, Earth Remote Sensing, Environmental Observation & Modeling, Seismology and ARM Climate Research. The aim of SciDB is to benefit all scientific applications dealing with large-scale complex scientific analysis and provide a way for scientists to understand data in far deeper and more natural ways.

We have also observed many works from distributed system community devoting to the adoption of cloud computing in scientific environments. Youseff et al. [72] establish a detailed ontology of the cloud, dissecting the cloud into five main layers: application layer, software environment layer, software infrastructure layer, software kernel layer and firmware/hardware layer. This ontology enables the scientific community to better understand the cloud technologies and design more efficient portals and gateways for the cloud. The Montage Comparison example [38] provides a detailed comparison between scientific workflow running in a local environment and running in a virtual environment. The experience shown in this paper gives the scientific community an idea what kinds of workflows are suitable to run on the cloud and what might be the cost if do so. [73] compared the performance of cloud to other platforms that are accessible to scientists. It also presented two main research directions in improving the cloud computing services for scientific computing, that is

to tune applications for virtualized resources and to optimize the application execution considering the cost-performance-security trade-off.

To ease the pressure of scientific community, people from distributed system community are working on simplifying the development process of scientific applications on the cloud. Aneka is a software platform for developing distributed applications on private and public clouds proposed in [74]. When implemented on the cloud, many scientific applications need to modify their original serial programs written in various programming models to parallel pattern. Since Aneka supports an extensible set of programming models, it can address a variety of different applications and thus offers a good opportunity for scientific applications to develop on the cloud with less effort.

National and Governmental Investment: Another opportunity lies in the construction of national cloud initiatives and the large amount of funding provided by major stakeholders, such as large user groups, vendors and governments, for cloud computing to achieve scientific and national objectives. With the utilization of commercially available technologies such as server virtualization, cloud computing is able to introduce capital cost savings to Information Technology (IT) infrastructure. Many nations have realized the importance of cloud computing to the modernization of IT. Cloud computing is a major feature of the US President's initiative to modernize IT and it's also taken as an important technology for the boost of Japan's economy by Japanese Government. Several national cloud initiatives have been announced, including the US Cloud Storefront, the UK G-Cloud and the Japanese Kasumigaseki.

The General Services Administration (GSA) of the US government is an agency that focuses on implementing projects that increase efficiencies and reduce operational cost by optimizing common services and solutions across enterprise and utilizing market innovations such as cloud computing services. In September 2009, the GSA's cloud storefront Apps.gov is launched by the Obama Administration. This online storefront enables federal agencies to efficiently and effectively acquire and purchase cloud computing services. Applications from desktop productivity toolsets to document management software are now available to buy through the online portal, which uses the software-as-a-service model to cut government IT purchasing costs.

The G-Cloud is an iterative programme of work to achieve government's commission to the adoption of cloud computing and delivering computing resources, which will deliver fundamental changes in the way the public sector procures and operates Information and Communication Technology (ICT). At present, still in its startup phase, the programme is resourced by existing departmental funding allocation whilst the dedicated business case (for 4.93 million pounds), to cover the ongoing staffing cost and development of the CloudStore is being developed, agreed and approved through the appropriate ministerial channels. The initial focus of G-Cloud is on introducing cloud ICT services into government departments, local authorities and the wider public sector. These services can then be reviewed and purchased through the CloudStore. At present there are 4 categories of services provided: Infrastructure, Soft-

ware, Platform and Specialist Services. The project savings in adopting cloud computing and re-using applications through the CloudStore can be broken down to G-Cloud & CloudStore and Data Centre Consolidation. It is estimated the savings of two kinds by year 2013, 2014 and 2015 will be £20m, £40m, £120m and £20m, £60m, £80m separately.

The Kasumigaseki Cloud initiated by the Japanese Government aims to establish a large cloud computing infrastructure to meet the resource requirements of the Government's IT systems and enable sharing to increase the utilization and efficiency of resources. A National Digital Archive will also be constructed to digitize government documents and recorded information and to improve the public access. The concept of Green Cloud Data Centers is used to construct the Kasumigaseki Cloud Data Center to reduce data center energy consumption by locating them in cold regions and increase the usage of green energy by utilizing wind and solar power.

6 CONCLUSIONS

eScience as a service is an emerging and promising service for scientific computing. In this survey, we develop a taxonomy and conduct a review on the current status of eScience services in the cloud with four kinds of sciences. Compared with the relatively mature grid infrastructure, the eScience tools and systems are in their early stage. We believe that eScience services will be boosted with more support from the cloud community and more investment and efforts from the science community. We call for the combined effort from both communities.

REFERENCES

- [1] G. Antoniu, A. Costan, B. D. Mota, B. Thirion, and R. Tudoran, "A-brain: Using the cloud to understand the impact of genetic variability on the brain," *ERCIM News*, vol. 2012, no. 89, 2012.
- [2] D. H. Hu, Y. Wang, and C.-L. Wang, "Betterlife 2.0: Large-scale social intelligence reasoning on cloud," in *CLOUDCOM '10*, 2010, pp. 529–536.
- [3] J. Ekanayake, S. Pallickara, and G. Fox, "Mapreduce for data intensive scientific analyses," in *ESCIENCE '08*, 2008, pp. 277–284.
- [4] A. Matsunaga, M. Tsugawa, and J. Fortes, "Cloudblast: Combining mapreduce and virtualization on distributed resources for bioinformatics applications," in *ESCIENCE '08*, 2008, pp. 222–229.
- [5] R. W. Zurek and L. J. Martin, "Gridpp: Development of the uk computing grid for particle physics," *Journal of Physics G: Nuclear and Particle Physics*, vol. 32, pp. 1 – 20, 2006.
- [6] J. Li, M. Humphrey, D. A. Agarwal, K. R. Jackson, C. van Ingen, and Y. Ryu, "eScience in the cloud: A modis satellite data reprojection and reduction pipeline in the windows azure platform," in *IPDPS'10*, 2010, pp. 1–10.
- [7] J. Li, M. Humphrey, Y.-W. Cheah, Y. Ryu, D. A. Agarwal, K. R. Jackson, and C. van Ingen, "Fault tolerance and scaling in e-science cloud applications: Observations from the continuing development of modisazure," in *eScience'10*, 2010, pp. 246–253.
- [8] M. Humphrey, Z. Hill, C. van Ingen, K. R. Jackson, and Y. Ryu, "Assessing the value of cloudbursting: A case study of satellite image processing on windows azure," in *eScience'11*, 2011, pp. 126–133.
- [9] "Dagman: A directed acyclic graph manager," Condor team, July 2005, <http://www.cs.wisc.edu/condor/dagman/>.
- [10] B. P. Rimal, E. Choi, and I. Lumb, "A taxonomy and survey of cloud computing systems," in *NCM '09*, 2009, pp. 44–51.
- [11] Association of Research Libraries, 2009, <http://www.arl.org/rtr/eresearch/escien/esciensurvey/surveyresearch.shtml>.
- [12] S. Bohle, "What is e-science and how should it be managed?" http://www.scilog.com/scientific_and_medical_libraries/what-is-e-science-and-how-should-it-be-managed/.

- [13] US Naval Research Laboratory, Monterey, Ca., <http://www.nrlmry.navy.mil/sec7532.htm>.
- [14] A. Newman, Y.-F. Li, and J. Hunter, "Scalable semantics - the silver lining of cloud computing," in *ESCIENCE '08*, 2008, pp. 111–118.
- [15] P. Watson, P. Lord, F. Gibson, P. Periorellis, and G. Pitsilis, "Cloud computing for e-science with carmen," in *2nd Iberian Grid Infrastructure Conference*, 2008, pp. 3–14.
- [16] "Top 500 supercomputer," 2010, <http://www.top500.org/lists/2010/11/>.
- [17] European Grid Infrastructure, <https://www.egi.eu/>.
- [18] University of California, 2012, <http://setiathome.berkeley.edu/>.
- [19] M. Peter and G. Timothy, "The nist definition of cloud computing," *National Institute of Standards and Technology*, October 7 2009.
- [20] I. Foster, Y. Zhao, I. Raicu, and S. Lu, "Cloud computing and grid computing 360-degree compared," in *GCE08*, 2008.
- [21] GRNET, <http://www.grnet.gr/>.
- [22] "Cloud-assisted real-time and large-scale monitoring and analysis for water quality," PDCC, NTU, 2011, <http://pdcc.ntu.edu.sg/camawq/home.html>.
- [23] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "A view of cloud computing," *Commun. ACM*, vol. 53, no. 4, pp. 50–58, Apr. 2010.
- [24] C. A. Lee, "A perspective on scientific cloud computing," in *HPDC '10*, 2010, pp. 451–459.
- [25] K. Keahey, "Cloud computing for science," in *SSDBM'09 (Keynote)*, 2009, pp. 1–1.
- [26] D. Oliveira, F. Baião, and M. Mattoso, "Towards a taxonomy for cloud computing from an e-science perspective," in *Cloud Computing: Principles, Systems and Applications*, 2010.
- [27] C. Hoffa, G. Mehta, T. Freeman, E. Deelman, K. Keahey, B. Berriman, and J. Good, "On the use of cloud computing for scientific workflows," in *ESCIENCE '08*, 2008, pp. 640–645.
- [28] "The us cloud storefront," 2009, <http://www.gsa.gov/portal/content/103758>.
- [29] "The uk g-cloud," 2009, <http://johnsuffolk.typepad.com/john-suffolk—government-cio/2009/06/government-cloud.html>.
- [30] "The kasumigaseki cloud concept," <http://www.cloudbook.net/japancloud-gov>.
- [31] V. Subramanian, L. Wang, E.-J. Lee, and P. Chen, "Rapid processing of synthetic seismograms using windows azure cloud," in *Proceedings of the 2010 IEEE Second International Conference on Cloud Computing Technology and Science*, ser. CLOUDCOM '10, 2010, pp. 193–200.
- [32] C. Evangelinos and C. N. Hill, "Cloud computing for parallel scientific HPC applications: Feasibility of running coupled Atmosphere-Ocean climate models on amazon's EC2," in *Cloud Computing and Its Applications*, 2008.
- [33] S. Nunez, B. Bethwaite, J. Brenes, G. Barrantes, J. Castro, E. Malavassi, and D. Abramson, "Ng-tephra: A massively parallel, nimrod/g-enabled volcanic simulation in the grid and the cloud," in *ESCIENCE '10*, 2010, pp. 129–136.
- [34] OpenNebula, <http://opennebula.org/users/users>.
- [35] EGI Federated Cloud Task Force, <https://www.egi.eu/infrastructure/cloud/>.
- [36] GAIA-Space, http://www.esa.int/Our_Activities/Space_Science/Gaia_overview.
- [37] Catania Science Gateway, <http://www.catania-science-gateways.it/>.
- [38] E. Deelman, G. Singh, M. Livny, B. Berriman, and J. Good, "The cost of doing science on the cloud: the montage example," in *SC '08*, 2008, pp. 50:1–50:12.
- [39] R. Curry, C. Kiddle, N. Markatchev, R. Simmonds, T. Tan, M. Arlitt, and B. Walker, "Facebook meets the virtualized enterprise," in *EDOC '08*, 2008, pp. 286–292.
- [40] N. Markatchev, R. Curry, C. Kiddle, A. Mirtchovski, R. Simmonds, and T. Tan, "A cloud-based interactive application service," in *E-SCIENCE '09*, 2009, pp. 102–109.
- [41] E. Deelman, D. Gannon, M. Shields, and I. Taylor, "Workflows and e-science: An overview of workflow system features and capabilities," 2008.
- [42] J.-S. Vöckler, G. Juve, E. Deelman, M. Rynge, and B. Berriman, "Experiences using cloud computing for a scientific workflow application," in *ScienceCloud '11*, 2011, pp. 15–24.
- [43] J. Wang and I. Altintas, "Early cloud experiences with the kepler scientific workflow system," *Procedia Computer Science*, vol. 9, no. 0, pp. 1630 – 1634, 2012.
- [44] M. Litzkow, M. Livny, and M. Mutka, "Condor - a hunter of idle workstations," in *ICDCS*, June 1988.
- [45] J. Dean and S. Ghemawat, "Mapreduce: simplified data processing on large clusters," *Commun. ACM*, vol. 51, no. 1, pp. 107–113, 2008.
- [46] S. Ghemawat, H. Gobioff, and S.-T. Leung, "The google file system," in *SOSP '03*, 2003, pp. 29–43.
- [47] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The hadoop distributed file system," in *MSST '10*, 2010, pp. 1–10.
- [48] L. Ramakrishnan, K. R. Jackson, S. Canon, S. Cholia, and J. Shalf, "Defining future platform requirements for e-science clouds," in *SoCC '10*, 2010, pp. 101–106.
- [49] OpenStack Swift, <https://swiftstack.com/openstack-swift/architecture/>.
- [50] G. DeCandia, D. Hastorun, M. Jampani, G. Kakulapati, A. Lakshman, A. Pilchin, S. Sivasubramanian, P. Voshall, and W. Vogels, "Dynamo: amazon's highly available key-value store," in *SOSP '07*, 2007, pp. 205–220.
- [51] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. E. Gruber, "Bigtable: a distributed storage system for structured data," in *OSDI '06*, 2006, pp. 15–15.
- [52] A. Lakshman and P. Malik, "Cassandra: A decentralized structured storage system," *SIGOPS Oper. Syst. Rev.*, vol. 44, no. 2, pp. 35–40, Apr. 2010.
- [53] R. Taylor, "An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics," *BMC bioinformatics*, vol. 11 Suppl 12, 2010.
- [54] P. G. Brown, "Overview of scidb: large scale array storage, processing and analysis," in *SIGMOD '10*, 2010, pp. 963–968.
- [55] SciDB Use Case, <http://scidb.org/use/>.
- [56] W. Wu, H. Zhang, Z. Li, and Y. Mao, "Creating a cloud-based life science gateway," *eScience*, vol. 0, pp. 55–61, 2011.
- [57] R. Alfieri, R. Cecchini, V. Ciaschini, and F. Spataro, "From gridmap-file to voms: managing authorization in a grid environment," *Future Generation Computer Systems*, vol. 21, pp. 549–558, 2005.
- [58] A. Nagavaram, G. Agrawal, M. A. Freitas, K. H. Telu, G. Mehta, R. G. Mayani, and E. Deelman, "A cloud-based dynamic workflow for mass spectrometry data analysis," *eScience*, pp. 47–54, 2011.
- [59] A. M. Thaufeeg, K. Bubendorfer, and K. Chard, "Collaborative eresearch in a social cloud," in *ESCIENCE '11*, 2011, pp. 224–231.
- [60] K. Chard, K. Bubendorfer, S. Caton, and O. Rana, "Social cloud computing: A vision for socially motivated resource sharing," *TSC*, vol. 99, no. PrePrints.
- [61] T. Dalman, T. Doernemann, E. Juhnke, M. Weitzel, M. Smith, W. Wiechert, K. Noh, and B. Freisleben, "Metabolic flux analysis in the cloud," in *ESCIENCE '10*, 2010, pp. 57–64.
- [62] J. C. Mudge, P. Chandrasekhar, G. S. Heinson, and S. Thiel, "Evolving inversion methods in geophysics with cloud computing - a case study of an escience collaboration," in *eScience*, 2011, pp. 119–125.
- [63] M. Malawski, G. Juve, E. Deelman, and J. Nabrzyski, "Cost- and deadline-constrained provisioning for scientific workflow ensembles in ias clouds," in *SC '12*, 2012, pp. 22:1–22:11.
- [64] E. S. Ogasawara, D. de Oliveira, P. Valduriez, J. Dias, F. Porto, and M. Mattoso, "An algebraic approach for data-centric scientific workflows," *PVLDB*, vol. 4, no. 12, pp. 1328–1339, 2011.
- [65] K. A. Ocana, D. de Oliveira, J. Dias, E. Ogasawara, and M. Mattoso, "Optimizing phylogenetic analysis using scihmm cloud-based scientific workflow," *eScience*, vol. 0, pp. 62–69, 2011.
- [66] H. Marty, S. Jacob, I. K. Kim, G. K. Michael, B. Jessica, and A. Michael, "Cloudn: A lightweight, end-to-end system for sharing distributed research data in the cloud," in *ESCIENCE '13*, 2013.
- [67] K. Chandrasekar, M. Pathirage, S. Wijeratne, C. Mattocks, and B. Plale, "Middleware alternatives for storm surge predictions in windows azure," in *ScienceCloud '12*, 2012, pp. 3–12.
- [68] D. de Oliveira, V. Viana, E. Ogasawara, K. Ocana, and M. Mattoso, "Dimensioning the virtual cluster for parallel scientific workflows in clouds," in *Science Cloud '13*, 2013, pp. 5–12.
- [69] Intel Hybrid Cloud Program, <http://www.intelhybridcloud.com/>.
- [70] The GoGrid Cloud Hosting, <http://www.gogrid.com/cloud-hosting/case-studies/>.
- [71] J. Becla and K.-T. Lim, "Report from the scidb workshop," *Data Science Journal*, vol. 7, no. September, pp. 88–95, 2008.
- [72] L. Youseff, M. Butrico, and D. Da Silva, "Toward a unified ontology of cloud computing," in *GCE '08*, 2008, pp. 1–10.
- [73] R. Iosup, S. Ostermann, N. Yigitbasi, R. Prodan, T. Fahringer, and D. Epema, "An early performance analysis of cloud computing services for scientific computing," *TU Delft, Tech. Rep., Dec 2008, [Online] Available*, 2008.
- [74] C. Vecchiola, S. Pandey, and R. Buyya, "High-performance cloud computing: A view of scientific applications," in *ISPAN '09*, 2009, pp. 4–16.